学校编码：10384　　分类号_____密级_____

学号：23020121152947　　UDC_____

厦门大学

硕 士 学 位 论 文

代价敏感决策树算法研究

Research on the Cost-sensitive Decision Tree Algorithm

叶林宝

指导教师姓名：冯少荣 副教授

专 业 名 称：计算机应用技术

论文提交日期：2015 年　月

论文答辩时间：2015 年　月

学位授予日期：2015 年　月

答辩委员会主席：_____

评　　阅　　人：_____

2015 年　月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行)》。

另外,该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。)

声明人（签名)：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘要

代价敏感学习是近几年数据挖掘领域的一个热门研究方向。基于代价敏感学习的分类算法的目标是进行分类时使得样例的误分类代价、属性检测代价等多种代价因素的总和最少。决策树作为一种经典的分类算法，其模型具有较好的可理解性、程序运行时较低的时间和空间复杂度、分类时较高的准确率等优点。正是由于决策树的诸多优点，近些年来不少学者尝试将决策树分类算法应用于代价敏感学习问题中。

现有的代价敏感决策树可以分为两类，一类是建立单一决策树模型来解决代价敏感分类问题，如 **EG2**、**PM**、**MinCost** 等；另一类是通过集成学习的方式来对样本进行代价敏感分类，如 **MetaCost**、**AdaBoost** 等。第一类算法的优点是执行效率非常高，而且具有较好的可理解性。第二类算法的优点是对样例进行分类时往往能得到更少的总代价，或者更高的分类准确度，但是这类算法执行时的时间和空间复杂度较第一类算法却高出了许多。这两类算法有一个共同的缺点，就是没有考虑到在分类过程中样例的某个属性的取值为离群值，或者连续型属性离散化过程中存在模糊性的情况对分类结果造成的糟糕的影响。

针对现有的第一类算法的特点，本文提出了相关的改进方法：（**1**）针对二分类和多类问题，本文分别提出了一种基于评分策略的代价敏感决策树，记为 **SECSDT** 和 **SECSDT_MC**。该算法在模型建立阶段充分考虑了代价因素和分类准确度因素之间的关系，分别对这两类因素进行评分来选择分裂属性。（**2**）在分类阶段，利用置信区间来识别样例中某属性的取值是否为离群值，或者离散化过程中可能出现模糊性的情况，然后利用多条决策路径进行分类的方式来得到最终的分类结果。（**3**）针对模型建立阶段中内部节点为选择合适的分裂属性而进行各种计算，造成建立模型的效率低下的问题，本文提出了相应的多线程版本的代价敏感决策树算法。实验结果表明，本文所提出的算法相比于已有的典型的代价敏感决策树算法具有更好的性能。

关键词： 决策树；代价；置信区间；评分

Abstract

Cost-sensitive learning is a hot research in the field of data mining in recent years. The target of classification algorithm based on cost-sensitive learning is to obtain the least of the sum of misclassification-cost, attributes-test-cost and other cost factors. Decision tree is a kind of classic classification algorithm, and it has the advantages of good comprehensibility, low time and space complexity, high classification accuracy and so on. As to so many advantages of decision tree, many researcher try to apply it to solve the cost-sensitive learning problems.

Existing cost-sensitive decision tree can be divided into two categories. The first kind is to create a single decision-tree model to solve the problem of cost-sensitive classification, such as EG2, PM, MinCost, etc. The other is combining multiple decision tree models to a new hybrid model, such as MetaCost, AdaBoost, etc. The first kind algorithm has the advantage of high execution efficiency and good comprehensibility. The other can obtain lesser total cost and higher classification accuracy than the first one, but it needs more time and space to produce the final decision tree model. These two kinds of algorithms have a common shortcoming, they both do not consider that some attribute value may be outliers or ambiguity exists in the process of continuous attributes discretization, and these can make bade effects on the classification results.

In view of the characteristics of the first kind of cost-sensitive decision tree algorithm, this thesis puts forward the relevant improving methods: (1) In the stage of establishing decision tree model under the condition of binary class and multiple class, this thesis puts forward the corresponding algorithm based on score-evaluation method. This algorithm fully considers the relationship of cost factors and classification accuracy, and then rate respectively for these two types of factors to select the most appropriate splitting attribute. (2) In the classification phase after establishing the cost-sensitive decision tree model, we

use the confidence interval to identify whether some attribute values are the outliers or ambiguity, and then choose more decision paths to get the final classification result. (3) It is inefficient in the stage of establishing the decision tree model, because it needs to do much various calculations in the internal nodes as to select the appropriate split attribute. This thesis also puts forward the corresponding algorithm of multiple-thread version. The experimental results show that the presented algorithms have better performance compared with the exiting typical cost-sensitive decision tree algorithm.

Key Words: Decision Tree; Cost; Confidence Interval; Score-evaluation

目录

第一章 绪论 1
1.1 研究背景与意义 1
1.2 国内外研究现状 2
1.3 本文研究内容 4
1.4 本文结构 5
第二章 代价敏感决策树算法基础知识 7
2.1 代价函数 7
2.2 代价敏感决策树模型的构建 9
2.3 分类方法 13
2.4 本章小结 16
第三章 基于评分策略的代价敏感决策树 17
3.1 基于评分策略的分裂属性选择方法 17
3.2 算法的收敛条件与类别判定 22
3.3 算法设计 23
3.4 多类问题的分裂属性选择方法 25
3.5 多线程版本的代价敏感决策树 27
3.6 本章小结 29
第四章 基于置信区间的分类方法 31
4.1 置信区间的产生 31
4.2 基于置信区间的分类方法算法设计 32
4.3 基于置信区间的分类方法实例 34
4.4 本章小结 36
第五章 实验设计与结果分析 37
5.1 实验设计 37

5.2 二类问题的代价敏感决策树算法性能 38
5.3 多类问题的代价敏感决策树算法性能 44
5.4 基于置信区间的分类方法 46
5.5 多线程版本的代价敏感决策树 50
5.6 本章小结 51
第六章 结论 53
6.1 总结 53
6.2 后续工作 54
参考文献 55
攻读硕士学位期间发表的论文 59
致谢..... 61
附件 1..... 63

Contents

Chapter 1 Introdoction 1

1.1 Background And Significance 1

1.2 Research Status 2

1.3 Research Content 4

1.4 Structure of the Thesis 5

Chapter 2 Preliminaries of Cost-sensitive Decision Tree 7

2.1 Cost Function 7

2.2 The Establishing of Cost-sensitive Decision Tree 9

2.3 Classification Method 13

2.4 Summary 16

Chapter 3 Cost-sensitive Decision Tree Base on Scoring-strategy 17

3.1 Splitting-attribute Selection Method Based on Scoring-strategy 17

3.2 The Convergence Condition And Categories Decide Method 22

3.3 Algorithm Design 23

3.4 Splitting-attribute Selection Method 25

3.5 Multiple Thread Version of Cost-sensitive Decision Tree 27

3.6 Summary 29

Chapter 4 Classification Method Based on Confidence Interval 31

4.1 Production of the Confidence Interval 31

4.2 Algorithm Design 32

4.3 An Instance of the New Clssification Method 34

4.4 Summary 36

Chapter 5 Experimental Study 37

5.1 Experimental Environment 37

5.2 The Performance In the Binary Clssification Problem 38

5.3 The Performance In the Multiple Class Clssification Problem 44

5.4 The Performance of New Classification Method 46
5.5 The Performance of Multiple Thread Version 50
5.6 Summary.....51
Chapter 6 Conclusions.....53
6.1 Conclusions.....53
6.2 Future Directions 54
References.....55
Papers..... 59
Acknowledgements 61
Attachment 1 63

第一章　绪论

1.1 研究背景与意义

决策树[1-3]是数据挖掘领域中的一种经典的分类算法。由于该算法相比于神经网络、贝叶斯等其它分类器在程序运行方面具有较少的时间和空间复杂度，构建出的模型易于理解，符合人类的逻辑思维，而且利用决策树构建的分类模型具有较高的分类准确度，因而被广泛用来解决各种分类问题。代价敏感学习作为数据挖掘领域 10 大最具挑战的问题之一，近几年来也备受关注[4]。

现有的决策树算法（例如 SPRINT[5]、SLIQ[6]、C4.5[7]）旨在尽可能获得最好的分类准确度，因此，在决策树构建过程中分裂属性的选择主要依托于基尼系数、信息熵、模糊规则等信息理论的方法。在模型构建阶段，这必然使得决策树的叶子节点倾向于将样例标记为训练集中样例数量较多的类别。这种模型并不适用于医疗诊断[8]、欺诈检测[9]、软件质量检测[10]、图像识别[11]等现实问题中，因为在现实分类问题中，常常涉及到多种代价因素（如误分类代价、属性检测代价等）[12]。例如在医疗诊断问题中，健康的人数往往大于患病的人数，然而将健康的人误诊断为患病的人造成的后果一般就是多花费了一些没必要的医疗费用，但是如果将患病的人误诊断为健康的人从而导致错过了应有的治疗，这种误诊造成的后果就不堪设想了，不是用金钱能够衡量的；又如在信用贷款问题中，将巨额金钱贷款给一个无法偿还的人造成的损失，远远大于拒绝将该金钱贷款给一个完全有能力偿还的人造成的损失。

基于决策树算法高效、准确度高、易于理解等优点，许多学者开始研究适用于代价敏感学习问题的决策树算法，在决策树构建过程中充分考虑问题涉及到的多种代价，例如上述医疗诊断领域中，将健康的人误分类为患者的代价、将患者误分类为健康的人的代价、诊断过程中检测各项数据需要付出的代价（例如血液检测、CT、B 超的费用）等。代价敏感决策树模型的目标就是在保证满足一定分类准确度的条件下，使得多种代价因素（包括类别的误分类代价，样例属性的检测代价等）的总和最小。

为了进一步提高分类准确度或减少分类总代价，不少学者提出遗传算法或者通过集成学习即组合多个决策树模型等方法来获得一种准确度更高、分类总代价更少的分类器，但是这些方式增加了计算的时间和空间复杂度，特别是随着属性维度的增加，计算的复杂度随之增加[13]。因此，根本的方法还是构建出一种能保证较高分类准确度的条件下，又能产生较少的分类总代价的决策树模型。本文提出的算法的代价因素主要指属性的检测代价和样例类别的误分类代价。

1.2 国内外研究现状

Kim 等人[14]通过理论分析和实验等方式证明了，代价敏感学习方法相比于非代价敏感学习方法能产生更少的分类总代价。目前，有关代价敏感决策树算法的研究主要有：(1)基于贪心方法建立单一的代价敏感决策树模型，例如 EG2[15]、PM[16-17]、MinCost[18]等；（2）利用非贪心的方法（遗传算法、boost 等）来建立多个决策树模型，然后对这些决策树模型的分类结果进行加权求和等方式来获得最终的分类结果，例如 MetaCost[19]、ICET[20]、TATA[21]、AdaBoost[22]等。

基于贪心方法建立单一的代价敏感决策树模型最大的优点就是计算复杂度低，效率高。这种类型的代价敏感决策树建立的关键点在于决策树模型中内部节点上分裂属性的选择方法，可以大致分为：

（1）使用基于信息理论（如信息熵、基尼系数、模糊集和隶属函数等）的方法选择分裂属性，如 C4.5[6]；

（2）综合考虑属性检测代价和基于信息理论的函数，如 EG2[15]；

（3）综合考虑误分类代价和属性检测代价，如 MinCost[18]；

（4）综合考虑误分类代价、属性检测代价和基于信息理论的函数，如 PM[16-17]。

事实上，构建性能较好的代价敏感决策树，本质上是使用最少的属性检测代价来尽可能地对样例进行准确分类，从而产生最少的误分类代价。Jan 等人[23]在论文中提出分类过程中分类准确度和分类总代价（误分类代价和属性检测代价的总和）这两方面往往不能同时得到最优解，因此这类算法的核心思想就是如何在满足较好的分类准确度的条件下，使得分类总代价尽可能的低。上述算法都是通